CENTRE FOR HEALTH
PROGRAM EVALUATION

WORKING PAPER 102

Life and Death: Theoretical and Practical Issues in Using Utility Instruments

Dr Graeme Hawthorne
Deputy Director, Program Evaluation Unit, Centre for Health Program Evaluation

Professor Jeff Richardson
Director, Health Economics Unit, Centre for Health Program Evaluation

Neil Day
Senior Research Fellow, Centre for Health Program Evaluation

Helen McNeil
Research Fellow, Centre for Health Program Evaluation

March, 2000
ISSN 1325 0663
ISBN 1 876662 15 8

CENTRE PROFILE

The Centre for Health Program Evaluation (CHPE) is a research and teaching organisation established in 1990 to:

- undertake academic and applied research into health programs, health systems and current policy issues;
- develop appropriate evaluation methodologies; and
- promote the teaching of health economics and health program evaluation, in order to increase the supply of trained specialists and to improve the level of understanding in the health community.

The Centre comprises two independent research units, the Health Economics Unit (HEU) which is part of the Faculty of Business and Economics at Monash University, and the Program Evaluation Unit (PEU) which is part of the Department of General Practice and Public Health at The University of Melbourne. The two units undertake their own individual work programs as well as collaborative research and teaching activities.

PUBLICATIONS

The views expressed in Centre publications are those of the author(s) and do not necessarily reflect the views of the Centre or its sponsors. Readers of publications are encouraged to contact the author(s) with comments, criticisms and suggestions.

A list of the Centre's papers is provided inside the back cover. Further information and copies of the papers may be obtained by contacting:

The Co-ordinator
Centre for Health Program Evaluation
PO Box 477
West Heidelberg Vic 3081, Australia
Telephone + 61 3 9496 4433/4434 **Facsimile** + 61 3 9496 4424
E-mail CHPE@BusEco.monash.edu.au

ACKNOWLEDGMENTS

This project was made possible by a project grant from the Victorian Health Promotion Foundation.

The Health Economics Unit of the CHPE is supported by Monash University.

The Program Evaluation Unit of the CHPE is supported by The University of Melbourne.

Both units obtain supplementary funding through national competitive grants and contract research.

The research described in this paper is made possible through the support of these bodies.

AUTHOR ACKNOWLEDGMENTS

We would like to acknowledge the generous cooperation and support of St Vincent's Hospital and the Austin and Repatriation Medical Centre. This thanks is extended to the ward and outpatient staff, including the nurses who were so helpful to our interviewers, often at times which were inconvenient for hospital routines.

Our thanks are especially extended to our interviewers who did such a terrific job of contacting and visiting the respondents: Sonia Barthelmebs, Kerith Culley, Janet Day, Marina Hawthorne, Ivan Marsetti, Jan Memery, Noelle Perry and Di Press. Their high level of professionalism and their willingness to work at times to suit the respondents all contributed to the success of this study.

We would also like to thank the respondents for completing interviews and questionnaires which were very long, and which were challenging because of the questions which probed those parts of their lives which were, for many respondents, associated with the losses which accompany illness.

This research was funded by the Victorian Health Promotion Foundation, to whom our thanks are extended. Without their generous support this research would not have been possible. Ethics approval was given by the Ethics Committees at Monash University, St Vincent's Hospital and the Austin and Repatriation Medical Centre.

Dr Hawthorne’s position is funded by the Victorian Consortium for Public Health, to whom our thanks are also extended.

Table of Contents

List of Tables

List of Figures

Background and Study Aims

The role of utility measurement is to provide an index which represents an estimate of the strength of preference for different health states. This index can be used to gauge preferences for the improvement in health-related quality of life (HRQoL) brought about by a health service or health program. In combination with epidemiological or clinical data on the expected duration of different health states, they may be used to calculate the number of quality-adjusted life years (QALYs), which is simply the product of the utility index and the time spent in the health state. This metric is the basis of cost utility analysis (CUA) which is increasingly being used by economists to evaluate health programs.

An important caveat is that for CUA utilities need to be obtained at the population or sample level; CUA is designed to evaluate programs, not individuals. A second important caveat is that the indices of utility must be correctly measured: they must be valid and reliable. As a number of different 'utility' values have commonly been obtained when different instruments — aptly described as 'synthetic indicators' by Nord [1] — are used to measure the same health state, the various instruments must differ in at least one of the following: what is measured, framing effects, weightings, or scoring mechanisms.

In the past two decades six utility instruments have been developed. Four are commonly used: the *15D*, *EQ5D* (or *EuroQo*l), *Health Utilities Index* (HUI2/3) and *Assessment of Quality of Life* index (AQoL); reviews can be found in Brazier *et al* or Richardson *et al*. [2, 3] Generally, instrument properties have been established through 'revealed validity'; i.e. *ad hoc* comparisons are carried out between the synthetic indicators' indices and criteria derived from clinical assessments or other HRQoL scores obtained on disease-specific non-utility instruments. Little work has been done on formal validation or on direct comparisons between utility instruments. Such work is necessary both to establish the meaning of what is measured and to give researchers information to make informed choices between the instruments.

This study investigated both these issues through comparing results obtained from the four instruments mentioned above, plus two commonly used generic non-utility instruments. The aims were (a) to examine overall utility scores; (b) to investigate the relationships between the instruments, (c) to explore the incremental utility scores within instruments and (d) to report how the instruments were similar/differed.

The Study Method:

Six HRQoL instruments were administered to a stratified sample of Victorian residents, selected to cover a very broad range of health conditions from those who were healthy through to those who were terminally ill. The strata were: (a) randomly selected community members weighted by socio-economic status to achieve representativeness of the Australian population; (b) outpatients attending two of Melbourne's largest public hospitals (the method used was random sampling within selected timeframes); and (c) inpatients from three Melbourne hospitals (purposive sampling was used within wards based on severity of condition).

The six instruments were the SF-36 and WHOQOL-Bref (generic health status instruments) and the AQoL, EQ5D, HUI3 and 15D (utility instruments). All instruments were scaled or scored as recommended by the developers. To avoid response bias instrument order was systematically rotated. This paper reports on the data analysis for the four utility instruments only.
A range of analyses were used, including scattergrams, correlations, analysis of variance and structural equation modelling.

The Utility Instruments

Each of the four instruments reported here consists of a 'descriptive system'; i.e. a series of item stems and responses which seek information about a concept or 'element' of the universe of HRQoL. Responses to these are then weighted and combined to produce the index. This section first describes each instrument, and then presents a brief discussion of some of the issues which arose in comparing them.

The Instruments

For the AQoL, the descriptive system comprises 15 items in 5 dimensions. Item responses are all ordinal scales with four levels per item. The dimensions are Illness, Independent Living, Social Relationships, Physical Senses and Psychological Wellbeing. [4] The utility weights were derived from an Australian population sample using time-trade off (TTO). During the calculation of the utility index, the Illness dimension score is not used. A multiplicative function is used to combine the remaining four dimensions into the utility index. [5]

The EQ5D (formerly the EuroQoL) consists of 5 items, each of which has 3 ordinal levels in the item responses. The items measure Mobility, Self-care, Usual Activities, Pain/Discomfort and Anxiety/Depression. The utility weights were obtained from a representative sample of the UK population, using the TTO. The utilities are computed using a regression model in which each item level is considered. [6]

The HUI3 comprises 15 items. The number of item responses varies between 4–6; again at an ordinal level. Of the 15 items, 12 are used in the utility score and form 8 'attributes'. These were constructed to be what the authors described as ‘within the skin’ attributes; that is, they focus upon disability and impairment rather than upon handicap. They are Vision, Hearing, Speech, Ambulation, Dexterity, Emotion, Cognition and Pain. The utility weights were derived using a visual analog rating scale (VAS), the values of which were transformed based on valuations obtained from the standard gamble. The weights reflect those of the Canadian population. As with the AQoL, the HUI3 uses a multiplicative model for combining the attributes into the index score. [7, 8]
The 15D consists of 15 items, and like the EQ5D each item represents a dimension. The 15D also focuses primarily on 'within the skin' dimensions, covering Mobility, Vision, Hearing, Breathing, Sleeping, Eating, Speech, Elimination, Usual Activities, Mental Function, Discomfort & Symptoms, Depression, Distress, Vitality and Sexual Function. The weights used were from the adult Finnish population and were elicited using rating scales. The data were transformed using Torrance's formulae [9] and the results interpreted as having utility properties. [10]

Some Issues

The four utility instruments reviewed here differ in virtually all respects; this makes direct comparability difficult.

First, the ‘perspective’ on HRQoL differs. The EQ5D offers a very plain functional perspective. The HUI3 (and the 15D) reflect a ‘within the skin’ perspective: that is items refer exclusively to impairment or disability: these instrument do not purport to measure handicap encountered in a social context, but impairment or disability to the contextless individual. The AQoL attempts to incorporate handicap and contains some questions probing the impact of impairment or disability upon a person’s life and social functioning.

Second, the descriptive systems differ in the dimensions included and the number of items in each dimension. This is shown in Figure 1 for the four utility instruments, and, by way of comparison, for the SF-36 also.

Figure 1 HRQoL CoverAge: Key Instruments

HRQoL dimensions	SF-36	AQoL	EuroQol	HUI-III	15D
Relative to the body					
Anxiety/Depression	***	*	*		**
Bodily care	*	*	*	*	
Cognitive ability				*	*
General health	******				
Memory				*	
Mobility	***	*	*	*	*
Pain	**	*	*	**	*
Physical ability/Vitality	*******			*	*
Rest and fatigue	**	*			*
Sensory functions		**		****	*****
Social expression					
Activities of daily living		*	*		*
Communication		*		**	*
Emotional fulfilment	**			**	
Family role		*			
Intimacy/Isolation		*			
Medical aids use		*			
Medical treatment		**			
Sexual relationships					*
Social function	**	*			
Work function	**				

Third, the different instrument designers adopted different methods for weighting the instruments. The 15D was weighted using a rating scale; the EQ5D and AQoL the time trade-off (TTO) method, and the HUI3 a rating scale which was then transformed into an estimate of standard gamble scores using a function fitted to selected health states for which both rating scale and standard gamble scores were obtained. In addition, the time period for which health states were to be endured also differed. For the AQoL and EQ5D the health state duration was specified as 10 years, while for the HUI3 the duration was a lifetime (defined as 60 years).

Fourth, the method of computing the utilities also varied. The 15D uses an additive model in which final disutility scores are a weighted average of the disutility for each item. The rating scale weights for the relative importance of each dimensions are re-scaled so that the weights sum to unity. The AQoL and the HUI3 use a multiplicative model in which a declining score on any dimension results in a fixed percentage decline in utility which remains after taking into account the disutility arising from the other dimensions. The EQ5D utilities are computed using a linear regression model derived from the econometric relationship between TTO scores for whole health states and the utility scores on each of the dimensions.

While at first it may appear that such diverse methods will inevitably result in very different estimates of health states utilities, this is not inevitable. It is possible to use quite dissimilar instruments to measure the same quantity and minimise error. For example, weight may be measured using either a spring or balance scale; distance, temperature and other physical quantities are commonly measured with different instruments employing different scales. Nevertheless, given the diversity of measurement strategies represented by the four utility instruments, disparate results would be unsurprising.

Results:

A. Participant Details

The response rates were 58% (n=396) for the community sample, 43% (n=334) for outpatients and 68% (n=266) for inpatients.

Details of participants are given in Table 1. This shows 50% of respondents were male, the mean age was 52 years, 75% were born in Australia, and 64% had attended either primary or high school. Forty-four percent were working in paid employment and 34% were retired. Sixty percent were married and 18% were single.

Table 1: Demographic Characteristics of Respondents

Gender	Male	488	50%
	Female	488	50%
Age	Mean (sd)	52.4	(18.0)
Birthplace	Australia	731	75%
	Other	245	25%
Education level (a)	Primary	116	12%
	High	488	52%
	TAFE/Trade	127	13%
	University	216	23%
Employment status	Fulltime	300	31%
	Parttime	126	13%
	Home duties	100	10%
	Student	30	3%
	Retired	328	34%
	Unemployed/Other	85	9%
Marital status	Single	175	18%
	Married/de facto	581	60%
	Separated/Divorced	105	11%
	Widowed	116	12%

Notes: The number of missing cases for any variable can be computed by subtracting the table entries from the base of 996.

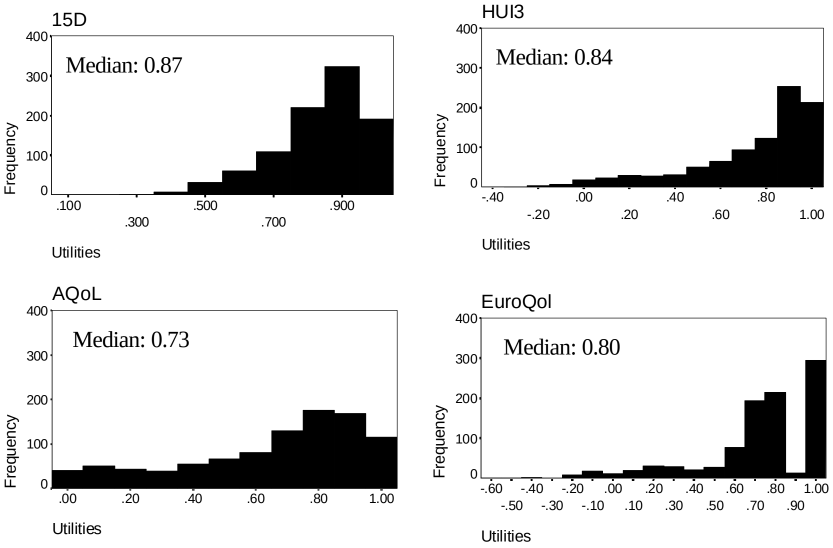
a = Highest level attended

B. Correlation Analyses

When basic psychometric tests were performed on each instrument, the most notable feature was in relation to missing data: the missing data rates were 2.8% (AQoL), 2.6% (EQ5D), 5.0% (HUI3) and 14.7% (15D). The high missing data rate for the 15D was due to 12.6% missing data on the 'sexual activity' question. Once these were modelled,[1] the rate was similar to that of the other instruments.

After missing data were modelled for all instruments, data distributions were examined and are shown in Figure 2. For the sake of comparison, all instruments have been placed on a common scale, where the boundaries were –0.60 through to +1.00 (representing health states worse than death and 'good' health respectively). The lower boundary of –0.60 in Figure 1 was set to be incrementally smaller than the all instrument lowest score (–0.59). This allows visual comparisons of the data distribution across the instruments. The theoretical lower boundaries, based on the scoring algorithms for each instrument, are –0.04 (AQoL), –0.59 (EQ5D), –0.36 (HUI3) and +0.11 (15D

Figure 2: Data Distribution



In economic evaluation the validity of utility measurement depends upon the validity of population summary statistics and, most commonly, the validity of mean scores. Mean instrument scores are given in Table 2. This shows that the highest utility values were obtained by the 15D and the lowest by the AQoL. As shown in Figure 2, however, these means may mislead.

[1] A regression formula is used for imputing missing data on the 15D; the AQoL uses the within dimension mean scores; no missing data procedures are provided with either the EQ5D or the HUI3.

Table 2 **Instrument Mean Scores**

	Mean	sd
AQoL	0.65	0.29
EuroQol	0.72	0.29
HUI3	0.74	0.27
15D	0.84	0.13

Figure 2 reveals that there are very different characteristics between the instruments' frequency distributions, even though the medians are similar. The differences occur in the distributions of the tails. For the 15D there is almost no tail, and scores were spread out over just 74% of the possible range. The HUI3 has a long tail of progressively declining values. For the EQ5D, there is a long tail with utility scores dropping sharply after 0.55, remaining relatively flat until 0.15, and then declining slowly into the negative range. There were very few scores in the vicinity of 0.90. The data distribution for the AQoL shows a more even spread of values over the scale range with a larger proportion of utilities in the range 0.00–0.50 than for the other instruments.

These findings lead to three important observations:

1. The variation in distributions suggests that the same individual may obtain different utility values depending on the instrument used. This implies there are differences in either the aspects of HRQoL measured, in the weights of the dimensions, or in the combination rules used to compute the utilities.

2. The histograms show that a relatively small proportion of respondents were in full health or were even close to full health as defined by any of the instruments, i.e. obtained utility values of, say, 0.90–1.00. *A priori*, it was expected that most of the community respondents would have had scores close to 'full health'; but this expectation was not fulfilled. For example, if 80% of community respondents had scores between 0.90–1.00 then 36% of all respondents should have fallen within this range. The results revealed that 19.5% of cases were within this region for the AQoL, 30% for the EQ5D, 40.5% for the 15D and 35% for the HUI3.

3. At the lower end of the utility scales, there were proportionally more respondents than expected. A utility value of 0.80 implies that a person would be willing to give up 20% of their life if it were possible to move to the health state described by the instrument as 'full health' (value: 1.00). For values less than 0.50 the score implies that the person is willing to give up more than 50% of their life. A priori, it would be expected that very few people would do this. However, the proportion of cases for each instrument falling below this value was 26.5% for the AQoL, 15.5% for the EQ5D, 2.2% for the 15D and 17.2% for the HUI3. These differences were just as pronounced at the lowest utility levels (what Furlong *et al* [8, pxii] referred to as the "pits" — where the evaluations provided utility scores worse than

death). For the AQoL the proportion of cases was 2.1%; for the EQ5D it was 3.8%; for the 15D[2] 0% and for the HUI3 2.1%.

Based upon the above, it might be expected there would be poor correlations between the different instruments, but this is not the case as shown in Table 3. The average Pearson correlation between the instruments was 0.75, suggesting that they were all providing highly related, although different, estimates of HRQoL. In relation to individual instruments, the lowest correlation was between the EQ5D and the HUI3 (r = 0.65); the highest was between the 15D and the AQoL (r = 0.82). To illustrate these differences, Figure 3 shows the scatterplot of AQoL and EQ5D scores. This reveals differences between estimates for many individuals.

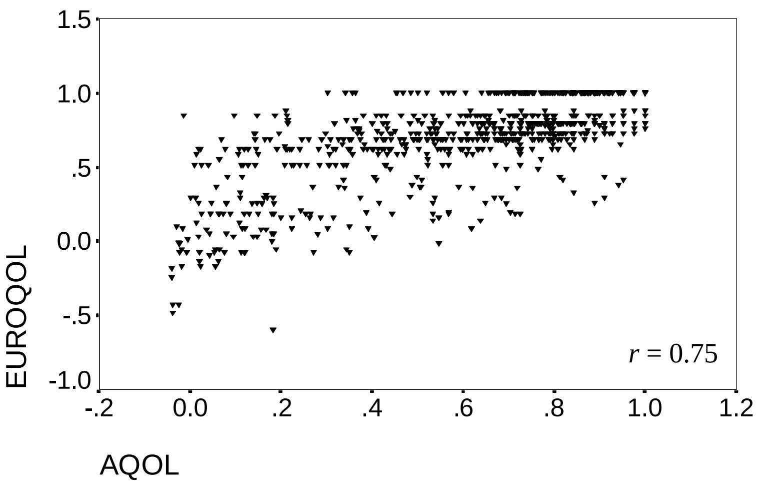
Table 3 Correlations Between Instruments

	AQoL	EuroQol	HUI3	
EuroQol	0.75			
HUI3	0.76	0.65		
15D	0.82	0.76	0.78	
Average *r*				0.75

As we have argued, these differences could be due to an instrument failure to measure a dimension, differences in the emphases given to each dimension in terms of the number of items, their content or the weights. It could also be due to individual response errors: individuals may misread or misunderstand items, and thus provide misleading or inconsistent information. Whatever the reason, the results shown in Table 2 and Figure 3 indicate that, at the present stage of its development, utility measurement at the individual level is a highly unreliable basis for prediction or decision-making.

[2] The lowest possible score for the 15D is +0.11 as reported above; i.e. it does not permit worse-than-death values.

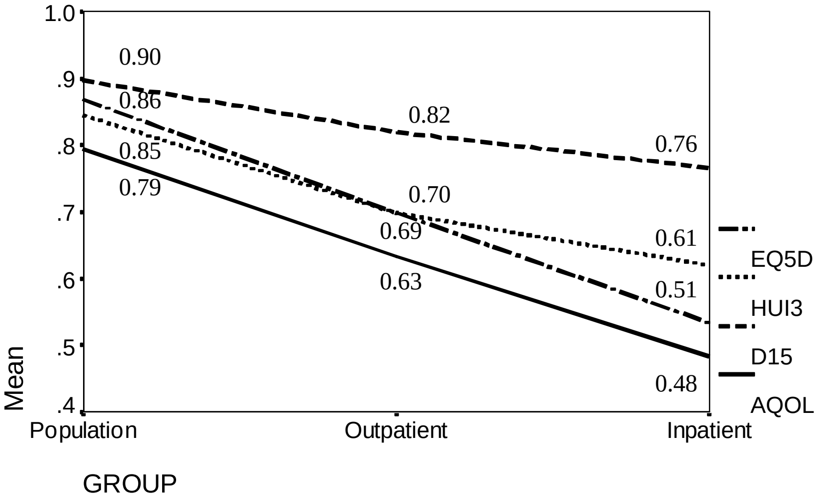
Figure 3 **Scatterplot: AQoL vs. EuroQol**



While there may exist a high correlation between instruments, the existence of a linear relationship between the different instrument scores is more important. In the context of CUA this linear relationship is of particular importance as the impact of a program is normally to increase health from one utility level to another. It is therefore important to know whether or not the different instruments predict, on average, the same incremental utility. Strictly this would require longitudinal (before and after) data. However it would be expected that cross-sectional data would reveal a linear relationship between instrument scores with a slope coefficient of unity, i.e. the difference in utility scores between the health states of two individuals would be expected to be the same when it is measured by different instruments.

As suggested by the scatterplot in Figure 3, the encouraging correlations shown in Table 2 mask a more complicated picture. The correlation analysis above assumes a constant relationship between the instruments for all respondents. A preliminary investigation into this assumption was undertaken by examining the relationship between mean utilities for each instrument according to respondent health status, by community, outpatient and inpatient samples. The results are presented in Figure 4.

Figure 4 **Discrimination by Status**



In this figure, each of the four lines was created by joining the mean utility values of the three patient groups for each of the four instruments. Consequently the gradient of each line was determined by the difference in the mean utility score for each patient group. As shown the gradients differ with the most marked discrepancy being between the HUI3 and the EQ5D. For the community and outpatient samples, these two instruments gave (effectively) identical utilities, yet the mean scores diverged for the inpatient sample by 20%. This indicates that the correlations between utility scores implied by Table 2 may not necessarily occur across the entire spectrum of health states. Thus, although a similar change in utility may be predicted by the EQ5D and HUI3 for community or outpatients, when the health state changes from best to poor health (e.g. among inpatients) this relationship may not hold true for a small change within the range of very poor health states.

C. Structural Analyses

The reason for these instruments were developed was to measure and place a value on HRQoL. This assumes that HRQoL exists as a construct, although it cannot be measured directly.[3] This implies each instrument comprises two components: the manifest variables (which respondents reply to) and a latent variable (the HRQoL that is being measured).

[3] It can exist in one of two ways. It is either a construct that is defined and created by the act of measurement or it exists independently of the measurement. If it is the former, then it is defined by the content of what is measured by each instrument and there is no necessary reason the different instrumens should correlate with each other. If it is the latter then the instruments should correlate with each other because they are attempting to measure a common construct. We believe it exists independently of our efforts to measure it.

This latent variable is inferred from the scoring algorithms of each instrument. These bring the disparate variables measured together into a single score; this score is assumed by the instrument developers to represent HRQOL. However, there are two problems: the content of utility instrument descriptive systems is contestable and the appropriate test of validity. The use of standard psychometric measures (e.g. exploratory factor analysis) may result in the exclusion of an item of importance to patients, but which is not part of a scale or dimension. [11] Psychometric tests are based on correlational analysis; if the correlations are low the item will load poorly and should be discarded. [12-14] The alternative approach is to accept the instrument developer's assertion of instrument content; i.e. an item is important based on the instrument developer's professional judgement (which may be informed by a variety of experts or patients etc.). Because there is no theory of validity that could be applied under the second scenario, we chose the first approach to convergent validity (i.e. using psychometric procedures to test that the items in the instruments are important).

In doing so we acknowledge that our use of structural equation modelling (SEM) implies a reflective latent indicator; *viz.* that HRQoL is driving the responses to the items within each instrument.[4] The psychometric approach is based on analyses of the extent to which items or dimensions correlate the latent variable, where the latent variable is defined by the items/dimensions themselves, as completed by respondents. Within this bounded modelling, there are two tests for validity: (a) that the items form a consistent group and (b) that there is homogeneity among the items. Where we have evidence of these we should obtain a congeneric model of good fit with substantial loadings on each of the indicators. It is our view that this congeneric validation is a necessary, although not sufficient, indicator of instrument validity.

It is expected that responses to the questions within each instrument (the observed HRQoL model) are driven by a higher order latent variable: HRQoL. Under this premise, examination of the relationship between the latent and observed models of HRQoL provides evidence of construct validity. Structural equation modelling analysis was used to investigate this. We used the weighted items/dimensions as shown in the figures, and not raw item scores to ensure that our models incorporated the item or dimension weights as closely as practicable.
Figures 5 through 8 show the results. For the EQ5D, HUI3 and 15D utility scores are derived directly from the instrument items; this implies a direct relationship between the latent HRQoL variable and the manifest variables. For the AQoL, there is a slightly different structure in that HRQoL is assumed to be a second order latent variable determining four[5] first order latent variables. These variables, in turn, affect responses to the manifest variables.

[4] This can be compared with a model in which there is a formative latent indicator, which is where the items within the instrument construct the latent variable.

[5] Although the AQoL has five dimension, the Illness dimension is not used in computation of the utility score.

Figure 5 **EQ5D structural Equation Analysis**

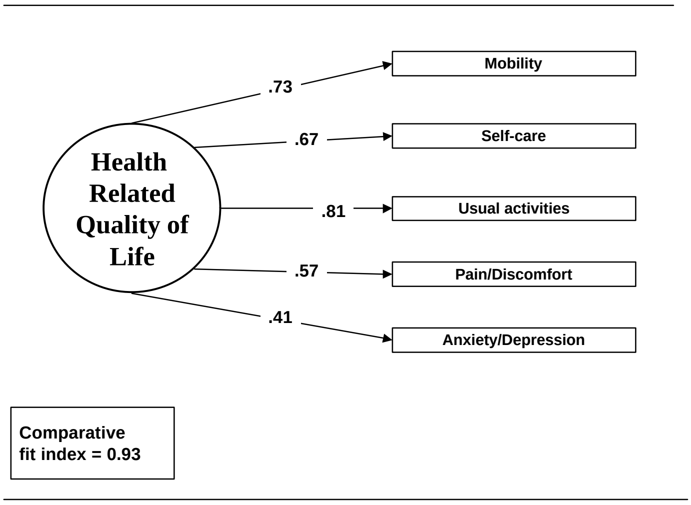


Figure 6 **HUI3 Structural Equation Analysis**

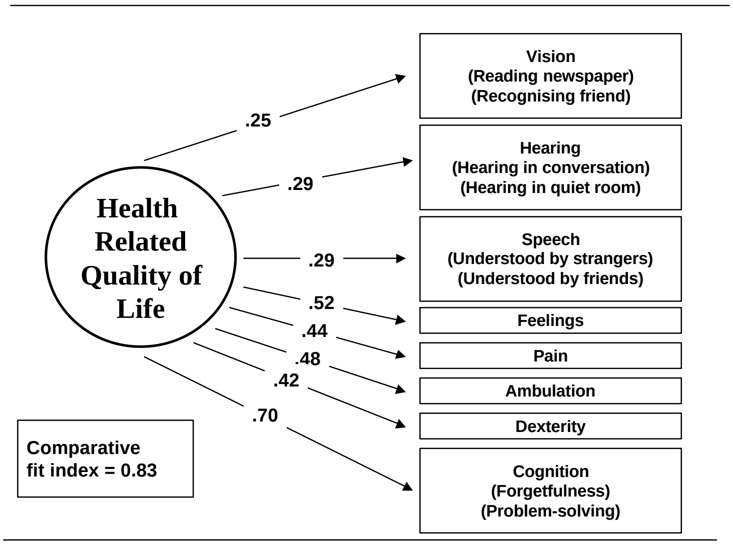

Health Related Quality of Life
.25
.29
.29
.52
.44
.48
.42
.70
Vision (Reading newspaper) (Recognising friend)
Hearing (Hearing in conversation) (Hearing in quiet room)
Speech (Understood by strangers) (Understood by friends)
Feelings
Pain
Ambulation
Dexterity
Cognition (Forgetfulness) (Problem-solving)
Comparative fit index = 0.83

Figure 7 15D Structural Equation Analysis

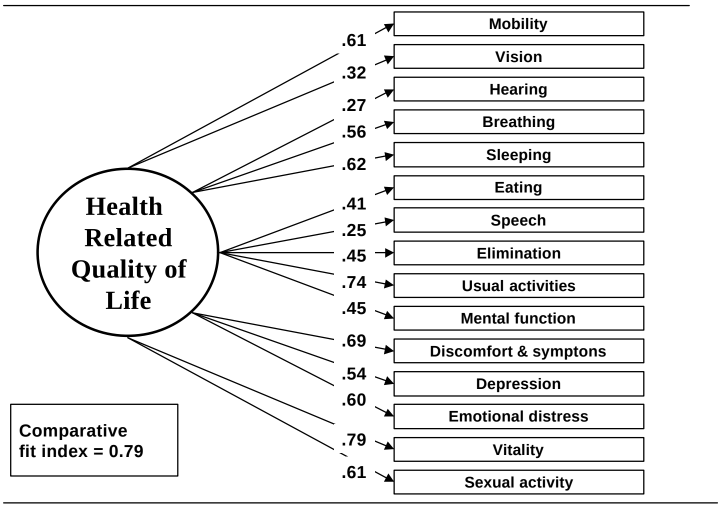
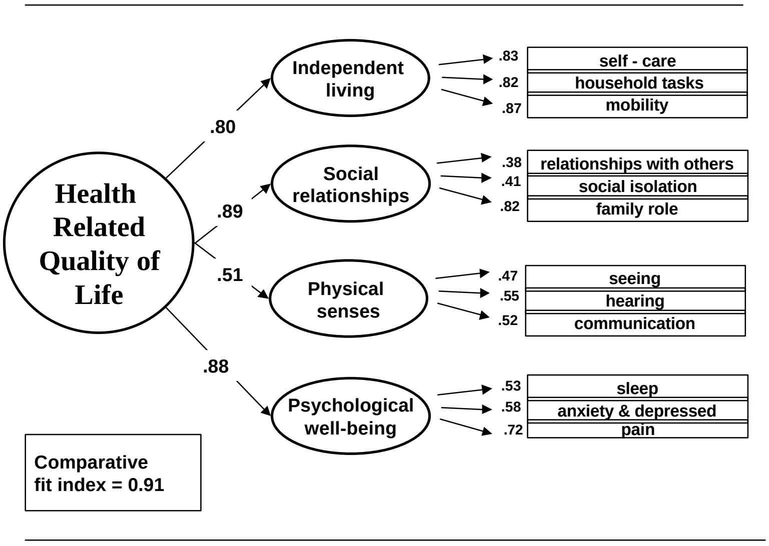


Figure 8 AQoL Structural Equation Analysis



The figures give the loadings associated with each manifest variable; squared loadings provide the proportion of variance in the item/dimension response which is explained by the latent variable. The overall goodness of fit statistic similarly shows the percentage of the variance which is explained by the model; this should be greater than 0.90. [13, 15, 16]
Of the four instruments, only two met this criteria: the AQoL (CFI: 0.91) and the EQ5D (CFI: 0.93). For the HUI3 the CFI was 0.83 and for the 15D it was 0.79. These findings suggest that the internal structures of the instruments differ, and that the empirically derived models for the HUI3 and 15D do not indicate a unitary concept of HRQoL.[6]

Of particular interest was the poor loadings obtained for those manifest variables measuring the physical senses: hearing, vision and speech/communication. For the AQoL these three items formed a single dimension which loaded in the model with a coefficient of 0.44, suggesting that 19% of variance within this dimension was explained by HRQoL. For both the HUI3 and the 15D, the average loading was 0.28 for these variables, indicating that 8% of variance on these measures was explained by HRQoL. These findings suggest that the physical senses are not as strongly related to HRQoL as some other dimensions; and it appears that the emphasis on these aspects of HRQoL within the HUI3 and 15D instruments may explain their poor CFIs.[7]

D. Comparative Analysis

The above findings from both the correlation and structural analyses suggest there are important differences between the four instruments. To investigate this further, with a view to explaining why these differences might occur, comparative analysis was undertaken.

Those respondents who obtained perfect utility scores on each instrument were identified, and their scores on the other instruments were examined. It was expected that the values on the comparator instruments would be clustered around '1.00'; the further the departure of utilities from '1.00' on the comparator instruments, the less the agreement between instruments. The results are shown in Figure 9. This shows that where a utility score of '1.00' was obtained on the AQoL or 15D, both the median scores and the distribution of other scores were also close to '1.00'. By way of contrast, where '1.00' was obtained on both the EQ5D and HUI3, the distribution of scores on the other instruments was broad. In the worst case, for example, when the EQ5D was set to '1.00' the median score on the AQoL was 0.90.[8] Although these findings were entirely consistent with the correlational analyses reported above, they should be interpreted cautiously. Although there is a notoriously high incidence of erratic individual scores in preference measurement, it must be remembered that the purpose of utility measurement is to obtain mean or median scores rather than individual scores. Subject to this caveat, the findings suggest that instrument selection may result in different utility values being obtained from cases with the same underlying

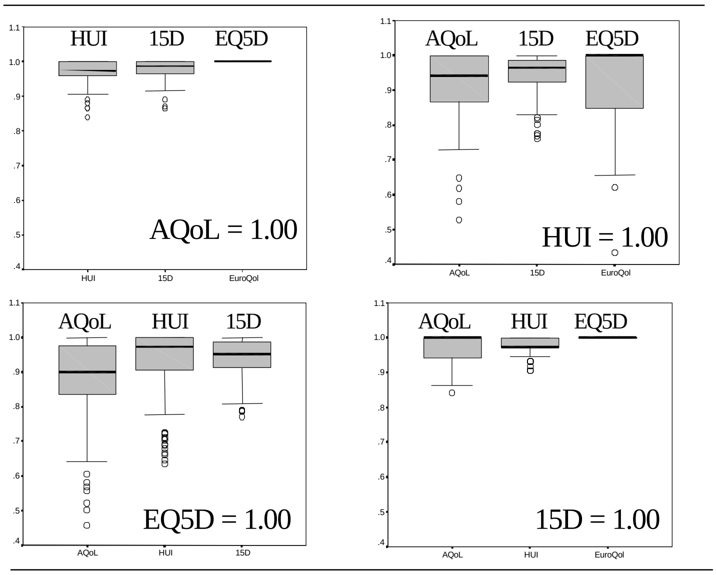
[6] We are not alone in this finding with respect to the HUI3. Richardson & Zumbo (in press) reported that the HUI3 primarily measured physical impairment and that this did not adequately represent the full gamut of health, including the physical, mental and social dimensions. Their exploratory factor analysis of the HUI3 revealed five dimensions: physical impairment, mental ill-health, mental wellbeing, general health impairment and social wellbeing. [17]

[7] These findings are consistent with those of Richardson & Zumbo (in press) [17]. They reported that physical impairment contributed 9% of the variance within the HUI3; a proportion they described as negligible.

[8] The lower median score of the AQoL is largely explained by the greater sensitivity of the instrument in the range of normal health and, consequently, the lower proportion of respondents obtaining 'full health' values (as shown in Figure 1).

health state. Whilst this was shown in the correlation analysis for those with poor health, the findings from the comparative analysis suggest it may also occur where respondents were in good health.

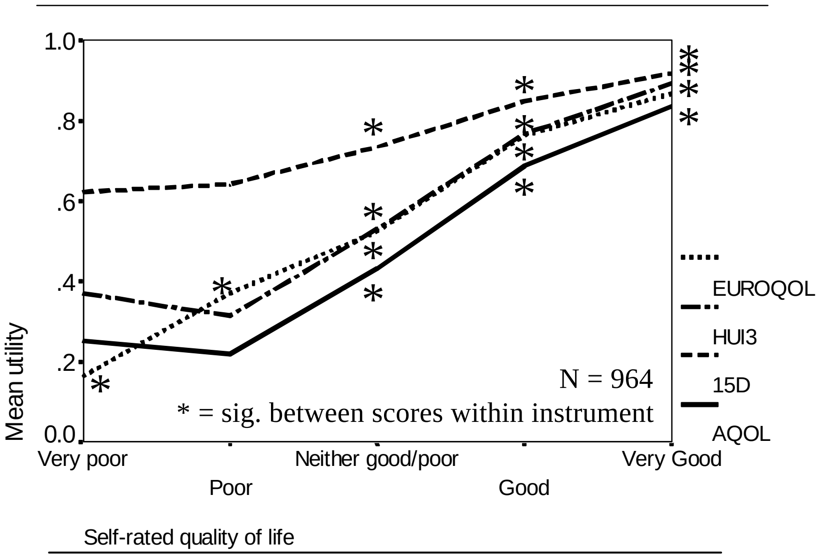
Figure 9 Effect of Different Instrument Selection



This suggested the need for further probing and this was undertaken by examining the degree to which respondents' own HRQoL evaluations converged with those of the MAU-instruments'. This was done in two ways.

First, as part of the WHOQOL, we asked respondents to rate their quality of life, using the question *How would you rate your quality of life*? Our expectation was that there would be a consistent and significant rank order increase in utility values for increases in the levels of self-rated quality of life (QoL). The results are shown in Figure 10. The figure demonstrates that the only instrument achieving this rank order property for all levels of self-rated QoL was the EQ5D: for each level the EQ5D utility values significantly increased in rank. For the other three instruments there were no significant differences between those who reported 'Very poor' and 'Poor' QoL, although the differences on each instrument between the other levels were significant in the postulated rank order. For the 15D the mean scores for those reporting 'Very poor' and 'Poor' QoL were virtually identical (0.62 vs. 0.64, Tukey HSD, >0.05, NS). However, for the HUI3 and the AQoL the mean utility scores were inconsistent: for the HUI3 the respective values were 0.37 and 0.31 (Tukey HSD, >0.05, NS). For the AQoL they were 0.25 and 0.21 (Tukey HSD, >0.05, NS). The Spearman correlations between the self-rated QoL and the obtained utility values were very ordinary: 0.59 for the AQoL and 15D respectively, 0.56 for the HUI3 and 0.50 for the EQ5D.

Figure 10: Self-rated QoL vs. Utility Instruments



Second, we administered both the MAU-instruments and a personal evaluation to a sub-sample of 121 cases. One of the difficulties we encountered in doing this was that many people were unwilling to trade on their own health. Of our sample, 68 cases (56%) obtained a value of '1.00'; yet the proportion of these 68 cases obtaining '0.90–1.00' on the utility instruments was 12% for the AQoL, 32% for the EQ5D, 40% for the 15D and 50% for the HUI3. Figure 11 presents the distribution of utility scores where the self-rated TTO was ‘1.00’ for all respondents with complete data (n=60). As would be expected from these findings, the correlations between the personal evaluations and the instrument utilities were 0.26 for the EQ5D, 0.28 for the HUI3, 0.36 for the AQoL and 0.45 for the 15D. Even allowing for the fact that individual preference scores usually have a poor correlation between instruments, these correlations are very low and suggest there are substantial differences between individual and population estimates.

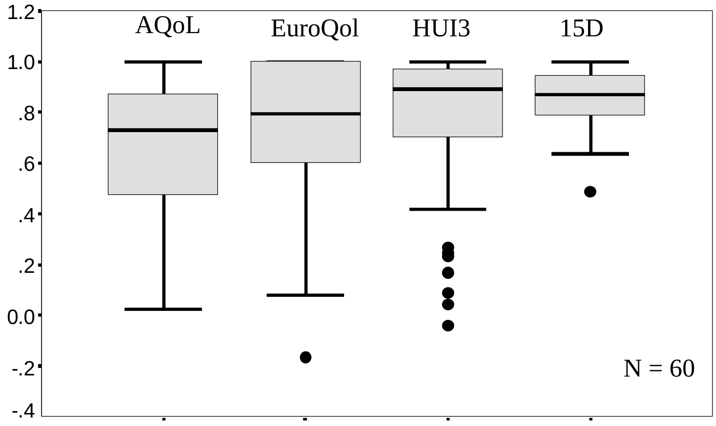
If MAU-instruments set out to reflect patient values then error-free self-rated utility scores ought to represent the 'gold standard'. The results reported in Figure 11 suggest that the four instruments in our survey would rate poorly against this gold standard.

Rightly or wrongly, however, there is broad consensus in the literature that community, and not patient values, should be incorporated into utility instruments. [2, 3] Consistent with this the weights in the instruments in this study represent those of the general population. Consequently the differences shown in Figure 11 between self-valuation and the MAU-evaluations may represent any of (a) the degree to which an individual has adapted to his/her health state [18], (b) their refusal to evaluate their own health state, (c) the extent to which personal views are systematically different to those of the population[9], or (d) to systematic error. Differences in national populations may also explain some of the variation in instrument results. If there are

[9] E.g. where the population evaluations reflect a 'shock horror' reaction to a given health state which it has not experienced.

cultural differences which affect utility values, the differences reported here may reflect this. In the field of hedonic psychology there are systematic national differences in life satisfaction [19]; and thus it would seem presumptuous to assume that preferences are universal. The AQoL uses Australian weights, the EQ5D British, the HUI3 Canadian and the 15D Finnish weights.

Figure 11 Self-rated TTO: 1.00 vs Utility Instruments



In seeking to explain these findings, it is pertinent to examine case studies of the differences between each instrument. Two cases are provided in Figures 12 and 13.

Case 1 (Figure 12) shows the impact of the broader coverage of the AQoL when compared with the 15D. The obtained utilities for this case were 0.14 for the AQoL and 0.55 for the 15D. As shown, the differences between these utilities were largely explained by the additional questions in the AQoL covering personal relationships and family role; neither of which are measured by the 15D. If these were set at '1.00' (i.e. no loss of HRQoL) in the AQoL (which is equivalent to the assumption that they were not measured) the AQoL utility rises to 0.49 — a value very close to that of the 15D.

Figure 12 Case 1

Health Dimension	AQoL	15d
Physical health and mobility	• Gets around home/community without difficulty • Has some difficulty focussing. • Hears normally.	• Walks normally has slight difficulty • Cannot read text; can see to walk • Hears normally • Shortness of breath on exertion • Eats normally • Serious bowel/bladder problems
Activities of daily living	• Needs no help with personal care • Or with household tasks	• Performs usual activities without difficulty
Bodily pain, General Health	• Suffers severe pain • Sleeps in short bursts only: is awake most of the night	• Severe physical discomfort/pain • Has great problems with sleeping. • Feels very weary
Social function	• **Has no close warm relationships** • Has friends and is not lonely • **Some parts of the family role affected by health.** • No difficulty communicating	• Speaks normally • Sexual activity almost impossible ?
Emotional and mental health	• Moderately anxious worried or depressed	• Feels extremely sad and anxious • Slight difficulties with thinking and memory
SF-36: FAIR HEALTH	**0.14 (0.49)**	**0.55**

In Case 2 (Figure 13), a similar analysis is shown using the HUI3 and the EQ5D. The obtained utility scores were 0.14 (HUI3) and 0.80 (EQ5D). The low HUI3 utility score was due to the low score obtained on questions measuring vision and hearing, neither of which are measured by the EQ5D. When these were set to '1.00', the HUI3 utility score rose to 0.74.

Figure 13 **Case 2**

Health Dimension	HUI-3	EuroQol
Physical health and mobility	• Walks without difficulty • Full use of hands and fingers • **Unable to see well even with glasses** • **Some hearing difficulty**	• No problems walking around ↔ ?
Activities of daily living	• Bathes, eats and dresses normally	• No problems with personal care • No problems performing usual activities
Bodily pain, General Health	• Moderate pain, occasionally disturbing normal activities • Health rated as fair	• Moderate pain or discomfort
Social function	• No problems with communicating	
Emotional and mental health	• Occassionally fretful, angry or depressed • Somewhat forgetful, but able to think clearly	• Not anxious or depressed
SF-36: AVERAGE HEALTH	**0.14 (0.74)**	**0.80**

These two case studies, and other similar results, suggest that the *content of what is measured* is a critical factor in determining utilities. Thus where an instruments' content does not cover the universe of HRQoL, the utility measurement is inflated and should be interpreted very cautiously. Conversely, where the universe of HRQoL is 'overmeasured', the utility estimate will be deflated due to double-counting. These findings indicate that where instrument content is incomplete the instrument will have poor content validity.

These findings need to be considered in light of the earlier findings from the structural equation analyses. The inclusion of a dimension within an instrument does not guarantee instrument validity, as this also depends upon the relative importance[10] of dimensions; it cannot be assumed that an instrument with broad coverage (e.g. the AQoL) is automatically valid. This point can be illustrated in the case of the HUI3. The SEM analyses showed that, at the study population level, the physical senses dimension explained the least variance in HRQoL. The current analysis has shown the physical senses play a critical role in determining the obtained utility scores because of their importance within the scoring algorithm. In Case 2, restoring full sight and hearing to a person who cannot see well and has some hearing difficulty would increase their HRQoL utility by 0.60 or by 330%.

[10] There are two different concepts of importance. On the one hand, importance can be defined as those health states with high weights; i.e. those health states which people will trade a lot on a TTO to avoid. But these may be uncommon and unique health states. On the other hand, there may be health states which are reported to be of importance from correlational analyses; i.e. they are pivotal in explaining people's HRQoL because they correlate highly with other items. Important is used here in the sense of possessing high weights.

Discussion

The findings from this validation study indicate that there are significant differences between the four utility instruments studied. These differences concern both the health states covered and the importance attributed to each. Analysis of content, modelling and utilities shows these instruments are not isomorphic: respondents with the same health states obtained different utilities depending upon which instrument was used. For this reason those who are contemplating using MAU instruments should be very careful in their instrument selection and interpretation.

The most important result was that the predicted change in utility is likely to depend upon the choice of instrument. For example, the evidence suggests that lower AQoL scores are often obtained because the AQoL includes health dimensions excluded by other instruments. In part the results are also a reflection of the different utility weights, which reflects both variation in modelling and perhaps the characteristics of the population from which the utility data were obtained.

Correlations between the scores produced by each instrument ranged from 0.65–0.82, but the data suggested these correlations systematically varied by respondents' health status. Structural equation modelling obtained comparative fit indices between 0.79–0.93. In two cases the evidence suggested the hypothetical model implicit in the scoring algorithm is incoherent; *viz*., there appears to be no single coherent concept underlying the manifest variables' data. Further, comparative analysis involving two different measures of self-report (QoL and TTO) suggested that none of the instruments provided utilities which were fully consistent with the expected results.

Conclusions

Overall, these findings suggest that none of the current state-of-the-art instruments possess sufficient validity for them to be used uncritically. The various instrument development teams need to issue careful comments about the limitations of their instrument and the way in which these results should be interpreted. The results suggest that before researchers select utility instruments they should thoroughly evaluate the properties of each instrument and, in particular, the content validity on the instrument in the health domain of relevance for their study. As a general rule it would be wise to include at least two generic instruments in a study. Decision-makers should be similarly aware of the possibility of bias due to the particular descriptive system or set of weights employed by the chosen instrument.

Despite these important reservations, at present these utility instruments are the best representations of consumer preferences that we have. Nord refers to them as 'synthetic indicators'. [1] The term is apt because it highlights the true nature of the indicators; i.e. particular constructs and not direct observations. The term also highlights the possibility of significant error. Notwithstanding, they should be included in economic and program evaluations, but with full awareness of their limitations.

References

1. Nord, E., A Review of Synthetic Health Indicators. Oslo, National Institute of Public Health for the OECD Directorate for Education, Employment, Labour and Social Affairs. 1997.
2. Brazier, J., Deverill, M., *et al.*, A review of the use of health status measures and economic evaluation. *Health Technology Assessment*, 1999. 3(9): p. 1-164.
3. Richardson, J., Olsen, J., *et al.*, The Measurement and Valuation of Utility-based Quality of Life. Melbourne, Centre for Health Program Evaluation. Working Paper 96. 1999.
4. Hawthorne, G., Richardson, J., & Osborne, R., The Assessment of Quality of Life (AQoL) Instrument: a psychometric measure of health related quality of life. *Quality of Life Research*, 1999. 8: p. 209-224.
5. Hawthorne, G., Richardson, J., *et al.*, Construction and Utility Scaling of the Assessment of Quality of Life (AQoL) Instrument. Melbourne, Centre for Health Program Evaluation. Working Paper 101. 2000.
6. Dolan, P., Gudex, C., *et al.*, Social Tariff for EUROQoL: Results from a UK General Population Survey. York, Centre for Health Economics, University of York. Discussion Paper 138. 1995.
7. Furlong, W.J., Torrance, G.W., & Freeny, D.H., Health Utilities index: Algorithm for determining Mark II/Mark III health status classification levels, health states and health state utility scores from 1992-10-20 self-administed health status questionnaire. Hamilton, McMaster University, Centre for Health Economics and Policy Analysis. 1996.
8. Furlong, W., Feeny, D., *et al.*, Multiplicative Multi-attribute Utility Function for the Health Utilities Index Mark 3 (HUI3) System: A Technical Report. Hamilton, McMaster University, Centre for Health Economics and Policy Analysis. Working Paper 98-11. 1998.
9. Torrance, G., Boyle, M., & Horwood, S., Application of multi-attribute theory to measure social preferences for health states. *Operations Research*, 1982. 30: p. 1043-1069.
10. Sintonen, H., The 15D measure of health-related quality of life: feasibility, reliability and validity of its valuation system. Melbourne, National Centre for Health Program Evaluation. Working Paper 42. 1995.
11. Juniper, E., Guyatt, G., & Jaeschke, R., *How to develop and validate a new health-related quality of life instrument*. In Quality of Life and Pharmacoeconomics. B. Spilker, Editor. 1996. Philadelphia: Lippincott-Raven Publishers.
12. Anastasi, A., Psychological Testing. 4th ed. 1976. New York: Macmillan Publishing.
13. Pedhazur, E.& Schmelkin, L., Measurement, Design and Analysis: An Integrated Approach. 1991. Hillsdale: Lawrence Erlbaum.
14. Streiner, D.& Norman, G., Health Measurement Scales: A Practical Guide to their Development and Use. 2nd Edition ed. 1995. Oxford: Oxford Medical Publications.
15. McArdle, J., Current directions in structural factor analysis. *Current Directions in Psychological Science*, 1996. 5(1): p. 11-18.
16. Bagozzi, R.& Heatherton, T., A general approach to representing multifaceted personality constructs: application to state self-esteem. *Structural Equation Modelling*, 1994. 1(1): p. 35-67.

17. Richardson, C.& Zumbo, B., A statistical examination of the Health Utility Index-Mark III as a summary measure of health status for a general population survey*. Social Indicators Research*, 2000. In press.

18. Frederick, S.& Loewenstein, G., *Hedonic adaptation*. In Well-being: The Foundations of Hedonic Psychology. D. Kahneman, E. Diener, and N. Schwarz, Editors. 1999. New York: Russell Sage Foundation.

19. Diener, E.& Suh, E., *National differences in subjective wellbeing*. In Well-being: The Foundations of Hedonic Psychology. D. Kahneman, E. Diener, and N. Schwarz, Editors. 1999. New York: Russell Sage Foundation.